Web からの情報抽出に基づく雑談的な対話の生成

吉野 幸一郎　　河原 達也
京都大学 情報学研究科

日々動的に変化する Web 上に存在するテキストから、述語項構造に着目した情報抽出を行うことで、雑談的な対話を生成する手法を提案する。対話の生成に情報抽出の枠組みを利用することで、「ユーザが何を知りたいか、何に興味を持っているか」に適合する情報を、より自然な形でユーザに提供することを目指す。具体的には、ユーザの知りたい情報が述語項構造に現れると考えて、それを解析することによって、ユーザの質問に回答したり、適合するものがなくても関連する情報を推薦する形で提供する。

1　はじめに

近年、Web 上の情報の爆発的な増加により、多くの知識や情報が Web 上に集積しており、それに伴い情報の検索、収集における Web の比重が高まっている。ただし、全ての情報要求が現状のキーワード型検索に適しているわけではなく、ユーザは漠然とした要求しか持たない場合も多い。そこで、ユーザの意図・嗜好を対話的に明確にしていくシステム（**情報コンシェルジェ**）の研究 [6] が行われている。この場合、対話システムの目的は情報検索のようなタスク達成ではなく、自然に雑談をしながら興味があることを知ることができる、ということにある。

従来実用化されてきた音声対話システムは、フライト情報案内 [1, 4] やバスの運行情報案内 [3] のように、特定のタスク遂行を目指し、関係データベース検索を扱ったものであった。これらの枠組みは、目的が明確である反面、データベースを定義する必要があり、構造化されていない大規模な Web データへの適用は難しい。それに対して、一般的な文書検索を対話的に行うシステムの研究 [9] も行われてきたが、既存の文書検索や質問応答システムのフロントエンドと位置づけられ、入力中のキーワードとその係り受け関係に着目しているものの、深い言語的解析や対話処理を伴わないものであった。その結果、対話中の文脈やユーザの意図に合わない不自然な応答が生成されることがあった。

また、情報コンシェルジェにおいてはユーザに対して情報推薦を行う機能が重要である。先行研究 [8] では、文書の中から特徴的な文章を抽出し提示していたが、必ずしも文脈やユーザの意図に沿っているとは限らなかった。これに対して本研究では、述語項構造の形で情報抽出を行うことで、Web 文書から的確な応答生成や情報推薦を行うことを目指す。さらに、ユーザとのインタラクションを利用したユーザモデルを定義することにより自然な情報推薦を行う。これによって、日々変化していく Web コンテンツから、自然な対話を生成することを目指す。

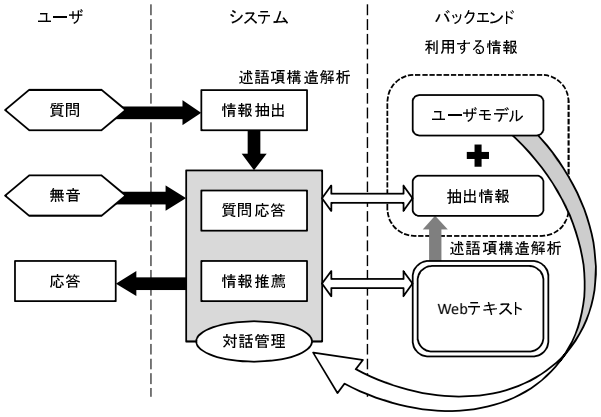

図 1　システムの概要

2　システムの概要

まず、本研究で提案する対話システムの概要について説明する。本システムは、Web に存在する Wikipedia、ニュースサイトなどの情報を利用して、ユーザの質問に答えながら対話を行うシステムである。今回は、プロ野球を主に扱うドメインとし、毎日新聞社のニュースサイトと Wikipedia のプロ野球関連記事を利用した。他のドメインにも順次拡張していく予定である。

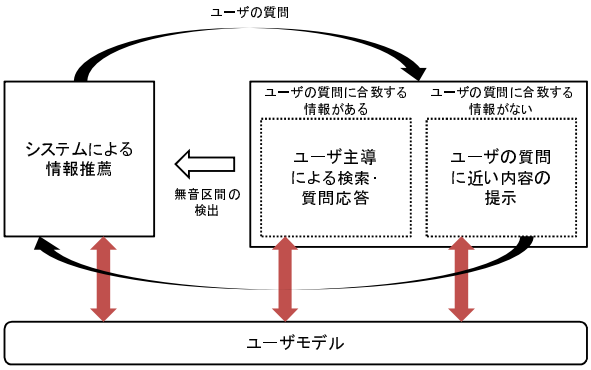

図 2　本システムの対話戦略

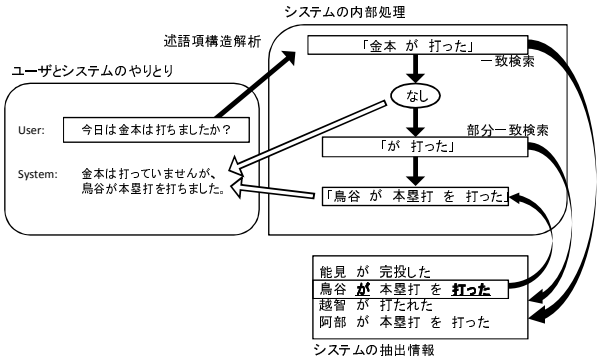

図 3　述語項構造の部分一致に基づく情報推薦

2.1　対話システムの概要

本システムの概要を**図 1** に示す。システムはあらかじめ Web から得たテキストを述語項構造解析し、情報抽出を行う。ユーザの質問に対しても同様に解析を行い、バックエンドで抽出された情報に対して、検索を行う。その結果、合致する情報があった場合は、該当部分の元テキストを基に応答を生成する。ユーザの質問に完全に合致する情報がなくても、「わかりません」と答えるのではなく、部分マッチングとユーザモデルを用いて質問の内容に近い情報推薦を行う。

2.2　対話戦略

本システムの対話戦略を**図 2** に示す。ユーザからシステムに対して質問があったとき、合致する情報があれば、その情報による応答を行う。完全に合致する情報がない場合、ユーザの質問に近い内容の検索を行い、その情報を提示する。例えば、**図 3** の「今日は金本は打ちましたか？」のようなユーザ発話がなされた場合、システムはユーザが意図する「金本が打ったかどうか」という質問に対する答えを持っていない。この場合、従来の検索型のシステムでは「情報がありません」「わかりません」のように答えるしかなかったが、述語項構造を用いた抽出情報の部分的なマッチングによって、「が　打った」に該当する情報をみつけて、その情報を提示する。これにより、ユーザとシステムの間の聞き返し、不自然な間がなくなり、情報を得るためのやりとりを増やすことができる。対話をしていく中で、ユーザもシステムも一定時間発話しない無音時間が検出されると、その直前に行った対話の内容を元に、情報推薦を行う。この結果、雑談的な対話を生成できることを期待している。

3　述語項構造の利用

3.1　述語項構造

文書中から情報を抽出し、対話の応答生成に利用するために述語項構造を用いた。述語項構造は

要素-格　用言

の形で文の意味を表現するもので

能見-が　抑えた　　(ガ格)
福原-を　打った　　(ヲ格)

のような例が挙げられる。述語項構造の利用は、自然言語理解において古くから行われてきたが、近年は、日本語における統計的な解析 [2, 5] も活発に行われている。本研究ではこの中で、**要素**を除いた**格＋用言**の形に着目した。文書検索において一致するキーワードがなくても、**格＋用言**の一致を探すことによって意味的に近い記事を検索し、提示する。すなわち、要素以外の部分一致検索に基づいて、**図 3** のような対話例を実現する。また、**図 3** に示されている

「鳥谷-が　本塁打-を　打った」

のように、複数の格＋用言のパターンを持つ文の場合は、部分的にマッチしたものも推薦できるように

「鳥谷-が　打った」
「本塁打-を　打った」

のようにそれぞれの格＋用言のパターンを取り出したものも保持しておく。

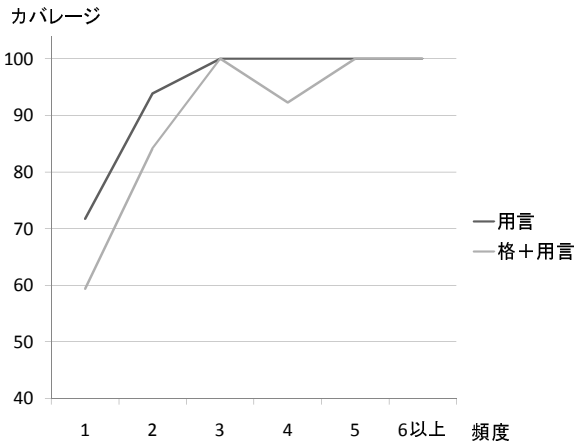

図 4　カバレージ

3.2　情報抽出のためのフィルタリング

新聞記事テキストに対して、KNP[*1]を用いて述語項構造解析を行い

要素-格　用言

の形で情報を抽出し保持する。この述語項構造解析の精度は、後述のテストセットに対して 89.3 %であった。質問応答や情報推薦においては、解析誤りの悪影響や、ドメインと無関係の情報を除く必要がある。また、野球のドメインにおける「放つ」と「打つ」のように、同じ意味を持つ用言などに対応する必要がある。そこで、ドメインの知識に基づいて有意味な格-用言パターンを抽出 (フィルタリング) すると同時に、同意表現の関連付けを行った。今回この作業を人手で行ったが、半自動化は今後の課題である。

毎日新聞記事データベース (CD-毎日新聞 2008 データ集) から、プロ野球に関連する記事 728 記事 (9910 文) を解析した。この結果、34228 個、12106 通りの格＋用言の組み合わせ、5855 通りの用言が抽出された。それからフィルタリング処理を行い、用言単位で約 2000 個となった。

3.3　カバレージ

毎日新聞社の web サイト[*2]から取得した 2009 年度クライマックスシリーズの新聞記事 309 文をテストセットとして評価を行った。

テストセットに出現する動詞、述語項（格＋用言）の頻度ごとにカバレージを計算した。この結果 (**図 4**) から、2 回以上出現するような表現は上記のようなパターン辞書を作ることで 85 %以上カバーできるものの、カバーできないものも多く存在することも確認された。全体の用言のカバレージは 75.2 %、格＋用言のカバレージは 63.7 %である。

4　ユーザモデルの定義

ユーザの意図や嗜好に沿った応答を生成する上で、ユーザモデルが有効である [7]。例えば、野球ドメインにおいて、嗜好は応援しているチームや選手にあると考えられる。そこで、チームと、打者・投手の区別に関して、モデル化した。各球団と関連付けるキーワードは、選手名鑑、Wikipedia から作成した。ただし、チーム、打者・投手の区別以外にも様々なモデル化をする必要があり、今後の課題といえる。

5　対話例

提案するシステムによる対話例を示す。

ユーザの質問に合致する回答がある場合の対話例を**図 5** に示す。最初のユーザの 2 つの質問に関しては該当する記事から応答を生成し、ユーザが話さなくなったところで無音を検出して情報推薦モードに入っている。

ユーザの質問に合致する回答がない場合の対話例を**図 6** に示す。元の記事にユーザの質問に完全一致する情報はなかったが、このユーザの興味が楽天の打者にあることを推測し、楽天の打者に関連する情報を探して、その情報を提示することで対話を生成している。

6　おわりに

Web 上のテキストから、雑談的な対話を生成するシステムを提案した。述語項構造解析を用いた情報抽出に基づいて頑健なマッチングを行うことで、質問応答や情報推薦を行った。これによって、ユーザの質問に合致した回答がなくても、ユーザの関心や焦点に沿った対話を行うことができるようになった。今後の課題としては、システムの評価を行うことと、他のドメインに対する適用を行い、汎用性を示すことが挙げられる。

[*1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
[*2] http://mainichi.jp/

User:	今日の野球の結果を教えて。
Sys:	中日がヤクルトに3－2で勝ち、1勝1敗としました。
User:	誰が勝ったの？
Sys:	中日は好投の吉見から浅尾、岩瀬と継投し、ヤクルト打線を抑え、逃げ切りました。
	（無音）
Sys:	吉見が今年タイトルを取ったのは知っていますか？
User:	知らない。
Sys:	吉見はセ･リーグの最多勝を取りました。
User:	パ･リーグは誰が取ったの？
	...

図 5　ユーザの質問に合致する回答がある場合

User:	今日は楽天はどうだったの？
Sys:	楽天が田中の好投で4－1とソフトバンクを破り、2連勝で第2ステージ進出しました。
User:	鉄平は打ったの？
Sys:	鉄平は打っていませんが、四回、1死一塁からのランエンドヒットで好機を広げ、2死一、三塁から中村真の左前適時打で先制しました。
User:	へー。他には？
Sys:	五回には2四球の走者を置いて山崎武が2試合連続の本塁打を放ちました。
	...

図 6　ユーザの質問に合致する回答がない場合

参考文献

[1] D.A.Dahl. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proc. ARPA Human Language Technology Workshop*, pp. 43–48, 1994.

[2] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[3] 安達史博, 河原達也, 奥乃博, 岡本隆志, 中嶋宏. Voicexml の動的生成に基づく自然言語音声対話システム. 情報処理学会研究報告, Vol. SLP-40-23, pp. 133–138, 2002.

[4] R.Pieraccini, E.Tzoukermann, Z.Gorelov, J-L.Gauvain, E.Levin, C.-H Lee, and J.G.Wilpon. A speech understanding system based on statistical representation of semantics. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 193–196, 1992.

[5] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A japanese predicate argument structure analysis using decision lists. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 523–532, October 2008.

[6] 河原達也, 川島宏彰, 平山高嗣, 松山隆司. 対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ. 情報処理, Vol. 49, No. 8, pp. 912–918, 2008.

[7] 河原達也, 荒木雅弘. 音声対話システム. オーム社, 2006.

[8] 翠輝久, 河原達也, 正司哲朗, 美濃導彦. 質問応答・情報推薦機能を備えた音声による情報案内システム. 情報処理, Vol. 48, No. 12, pp. 3602–3611, 2007.

[9] 翠輝久, 駒谷和範, 清田陽司, 河原達也, 木戸冬子. 音声対話による大規模知識ベース検索システム -音声版ダイアログナビ-. 情報処理学会研究報告, Vol. SLP-52-4, pp. 21–26, 2004.